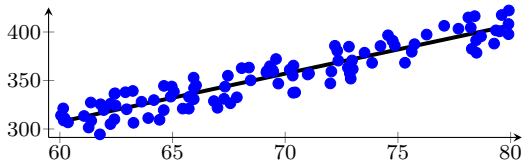
4.1: Visualizing Variability with a Scatterplot

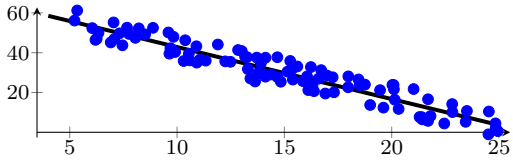
Definition.
A **scatterplot** is used when examining the relationship between two *numerical* variables where each point represents an observation. With scatterplots, we examine the

- **trend**: general tendency of the scatter plot going from left to right
- **strength**: strong associations have little vertical variation
- **shape**: is the scatterplot linear or nonlinear?

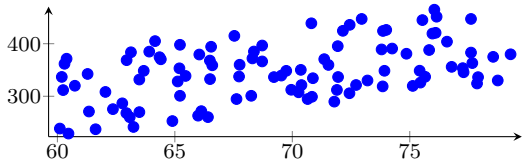
Positive trend



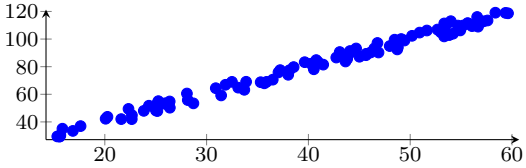
Negative trend



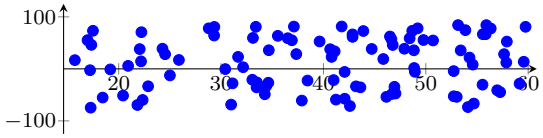
Weak association



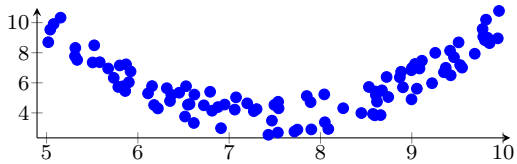
Strong association



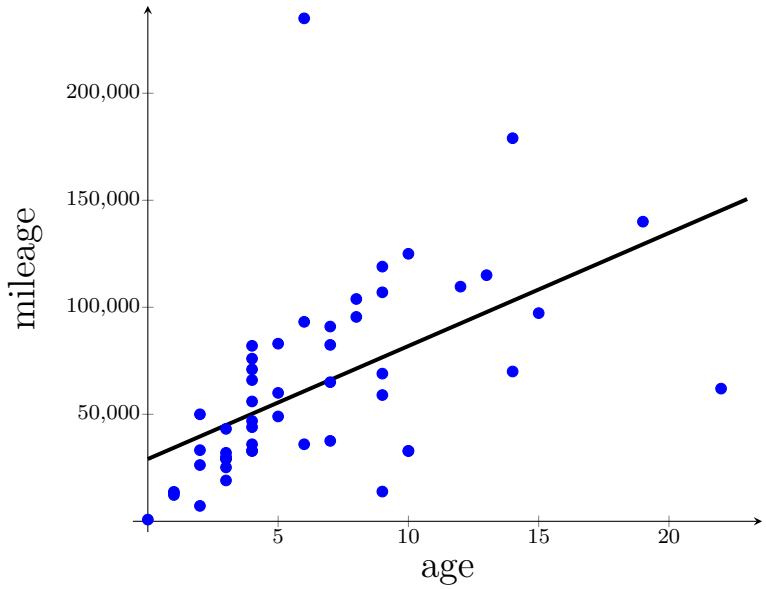
No trend



Quadratic/nonlinear shape



Example. The scatterplot below shows the age and corresponding mileage for a sample of used cars.



What is the association between the variables? Identify the trend, its shape, and how strong the relationship is.

Trend: Positive trend
Shape: Linear
Strength: Medium strength

4.2: Measuring Strength of Association with Correlation

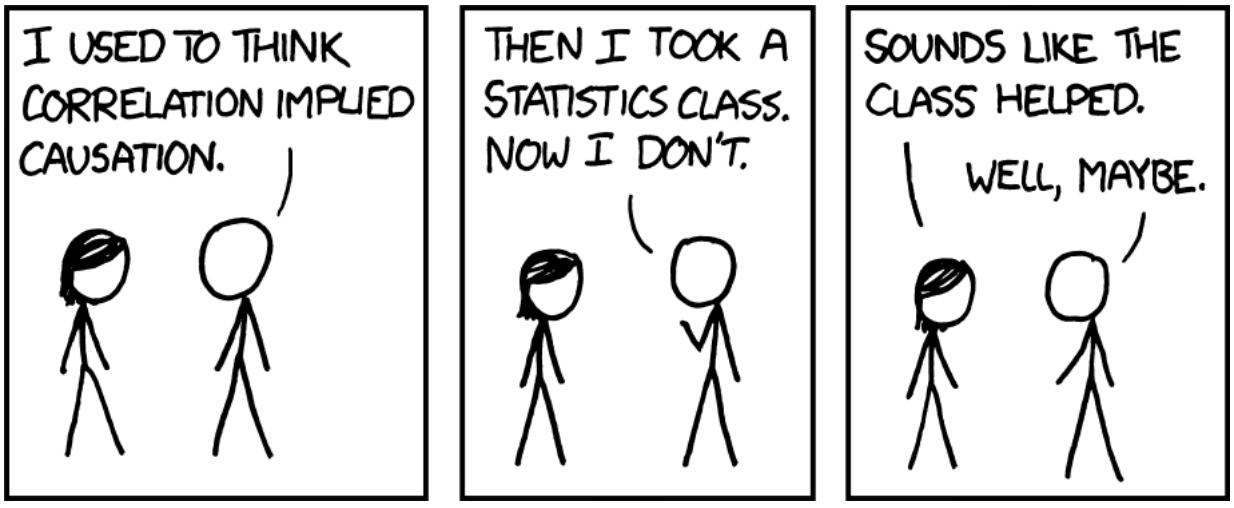
Definition.
The **correlation coefficient** is a number that measures the strength of the linear association between two numerical variables. The correlation coefficient is between -1 and 1:

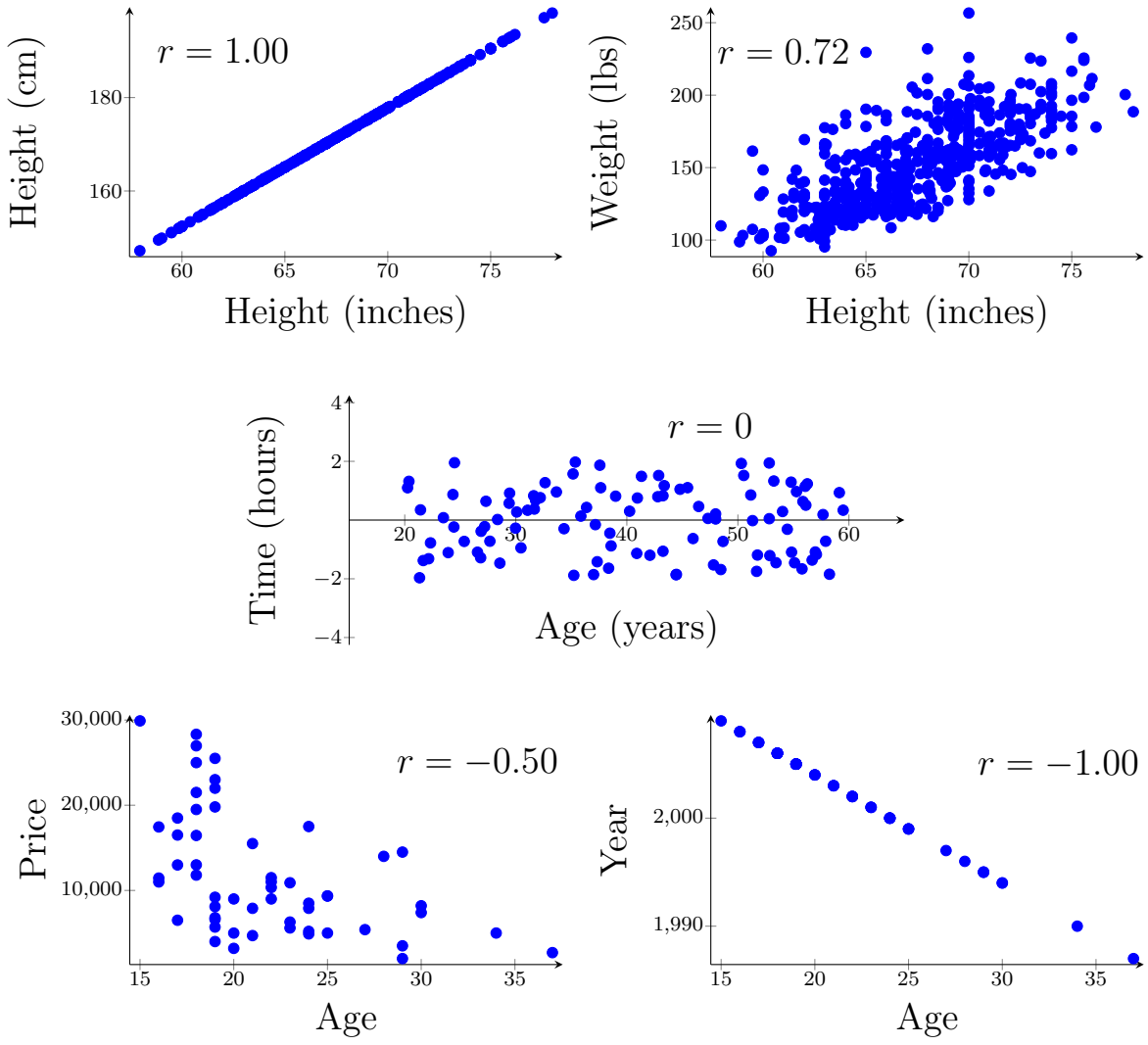
$1 \rightarrow$	Strong association and positive trend
$0 \rightarrow$	Weak or no association
$-1 \rightarrow$	Strong association and negative trend

The correlation coefficient only makes sense if the trend is linear and both variables are numerical!!

Note: *Correlation does not mean causation!*

Take a few minutes to look at the graphs at this link: Spurious correlations




$r = 1.00$
Height (cm)
180
160
60
65
70
75
Height (inches)
$r = 0.72$
Weight (lbs)
250
200
150
100
60
65
70
75
Height (inches)
$r = 0$
Time (hours)
4
2
−2
−4
20
30
40
50
60
Age (years)
$r = -0.50$
Price
30,000
20,000
10,000
15
20
25
30
35
Age
$r = -1.00$
Year
2,000
1,990
15
20
25
30
35
Age

Definition.
The formula for the correlation coefficient between two variables x and y is

$$r = \frac{\sum z_x z_y}{n-1}$$

where z_x and z_y are the z scores for each entry in the x and y lists.

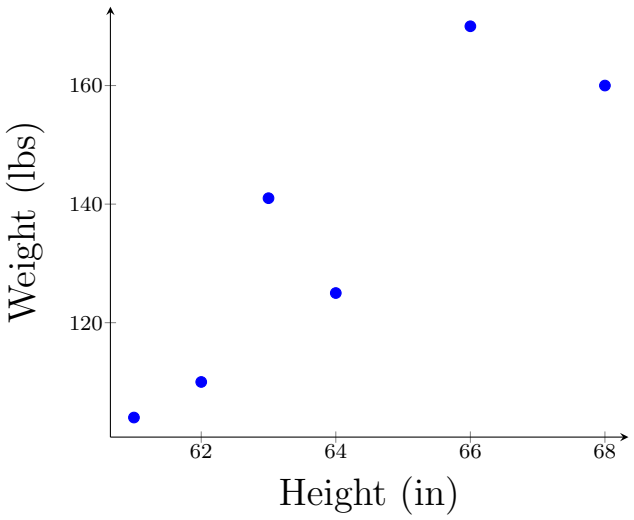
$$z_x = \frac{x - \overline{x}}{s_x} \qquad z_y = \frac{y - \overline{y}}{s_y}$$

Example. Below are the heights and weights of six women:

Heights	61	62	63	64	66	68
Weights	104	110	141	125	170	160

Compute the correlation coefficient by hand. Then, graph the scatterplot and compute the correlation coefficient using StatCrunch.

	x	y	zx	zy	zxzy
	61	104	-1.1504	-1.1598	1.3343
	62	110	-0.7670	-0.9353	0.7174
	63	141	-0.3835	0.2245	-0.0861
	64	125	0.0000	-0.3741	0.0000
	66	170	0.7670	1.3095	1.0043
	68	160	1.5339	0.9353	1.4347
				sum	4.4047
mean	64	135		sum/(n-1)	0.8809
std dev	2.60768	26.7283			



Understanding the Correlation Coefficient:

- Changing the order of the variables does not change r
- Adding a constant or multiplying by a positive constant does not affect r
- The correlation coefficient is unitless
- None of this makes any sense if the relationship between the variables is not linear!